Multi-masks Generation for Increasing Robustness of Dense Direct Methods

Ziming Liu [1] and Ezio Malis [1] and Philippe Martinet [1]

***Abstract*— In this paper, we address the problem of increasing the precision of dense direct stereo visual odometry methods. Dense methods need a dense depth map to generate warped images (virtual views) that will match with reference images if the estimated pose is good. Previous works have shown that generating the depth map by machine learning methods leads to very good odometry results. However, machine learning methods generate hallucinated depths even in areas where it is impossible to estimate the depth due to several reasons, like occlusions, homogeneous areas, etc. Generally, this produces wrong depth estimation that leads to errors in odometry estimation. To avoid this problem, we propose a new approach to generate multiple masks that will be combined to discard wrong pixels and therefore increase the accuracy of visual odometry. Our key contribution is to use the multiple masks not only in the odometry computation but also to improve the learning of the neural network for depth map generation. Experiments on several datasets show that masked dense direct stereo visual odometry provides much more accurate results than previous approaches in the literature.**

I. INTRODUCTION

The dense direct visual odometry method (DDM) is one of the most popular approaches for visual odometry [1], [2], [3], [4]. In DDM method, a cost (or loss) function is computed from a target image and an image warped with an estimation of the pose and of the depth map, see Fig. 1. The optimal pose is found by optimizing the cost function. This scheme is very similar to the deep learning-based depth estimation network proposed in [5], [6], [7], [2] where image loss can also be used for training the network.

However, when learning depth maps with networks we can not avoid the problem of hallucinating depths in some part of the image, such as areas where the texture is homogeneous, or areas with stereo occlusions or temporal occlusions. Typically, incorrectly estimated depth result in incorrect image warping losses and reduced performance in visual odometry. These problems limit the accuracy of the depth map estimation networks and the DDM visual odometry methods.

The methods using the image-based loss has achieved great success in many tasks, mainly in deep learning(DL)-based depth estimation, and visual odometry [8], [1]. For most unsupervised depth estimation networks, photometric image warping loss has shown good performance without any ground truth depth annotation [8], [6]. For visual odometry, the DDM does not need to compute feature detection and feature matching/tracking. This not only reduces the possible errors from feature detection, but also saves the time for constructing feature descriptors [1].

Recent state-of-the-art unsupervised depth estimation works have widely used the image-based loss optimization. Meanwhile, many of these works have shown the importance of applying masks on the loss during optimization. The proposed masking methods mainly focus on solving the problems of removing hallucinated depth areas. A group of them choose to define geometric rules to obtain masks. [6], [9], [10], [11]. However, these masks generation methods highly rely on accurate disparity predictions, and most of them only consider temporal context. There are also some works using deep network to generate masks [7], [12]. These approaches need more computations and parameters. Another problem for them is the ground truth mask annotations for the network training or validation can not be obtained.

Similarly, the state-of-the-art dense direct method visual odometry methods also use different masks, such as certainty and rigid masks [3], [4]. They all show the positive effect of applying masks on the DDM. But these works only partially solve the occlusion or the homogeneous texture problems of the DDM.

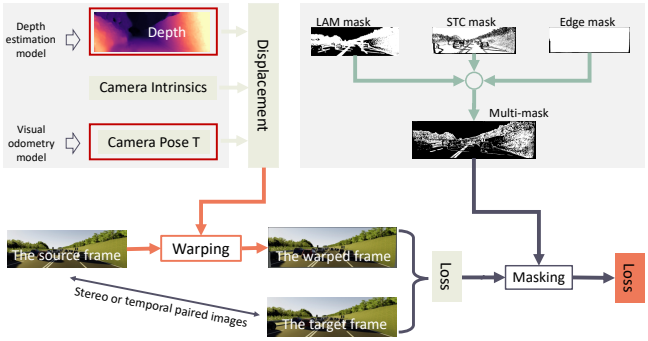


Fig. 1. The image-based loss with the proposed multi-mask system. This loss is measured comparing the warped frame and the target frame. The multi-mask contains three kinds of masks, used for masking the loss. For the depth estimation stage, the predicted depth map is optimized. For the visual odometry, the relative camera pose from the DDM is optimized.

In this paper, a new multi-mask approach is proposed to increase the accuracy of the image-based loss. This new method is called Masked DDM. The Masked DDM contains two parts, one part is the depth estimation network trained with the masked image-based loss. Another part is the dense direct visual odometry module which is optimized with the masked image-based loss at the inference stage. Firstly, to reduce the effect of the occlusion areas, we consider the stereo-temporal relation of the stereo warping and temporal warping

[1] All authors are with ACENTAURI team at INRIA (Sophia Antipolis, France), and 3IA Côte d'Azur, Université Côte d'Azur (Nice, France). {first name}.{name}@inria.fr

in a same view. For the non-occluded pixels of the view $\{camera\ i, time\ t\}$, these pixels can be warped by the stereo warping of another stereo camera or by the temporal warping of the adjacent frame. According to the above motivation, a Stereo-Temporal Consistency (STC) occlusion mask is proposed. To be robust for the brightness discrepancies of different camera views, a ZNCC measurement is introduced to measure the error in the mask on the local image patch level. Secondly, the homogeneous texture areas, such as the sky, also result in the hallucinated depths. Therefore, a Local Average Max (LAM) homogeneous texture mask is proposed to solve that. Finally, the STC mask, the LAM mask and the edge mask [11] construct the multi-mask system.

The main contributions of this paper are (i) a new Masked DDM visual odometry method, (ii) a stereo-temporal consistency mask for filtering the occlusion pixels, and the ZNCC measurement is firstly introduced for solving the brightness discrepancies problem of different views, (iii) another simple but efficient mask to find the homogeneous texture pixels.

This paper is organized as follows. Section II describes the details of the proposed method. Section III gives the experimental results and analysis. Section IV presents the conclusions and future research directions.

II. Masked DDM

Here, we first introduce the overall Masked DDM, then describe the details of the proposed multi-mask system, including the STC mask, the LAM mask, and the edge mask.

A. The Overall Masked DDM

The masked DDM is based on a hybrid visual odometry method [2], and combined with the proposed multi-mask system. It has two main parts, including the depth estimation network and the pose estimation module. This depth estimation network is trained using the image warping loss on the calibrated stereo or temporal image pairs, without ground truth depth annotations [5], [6], [2]. This loss will be combined with the multi-mask system in the training stage. For the inference stage, the predicted depths are fed into the pose estimation module, which is optimized online with the image warping loss combined with the multi-mask.

B. Stereo-temporal Consistency (STC) Occlusion Mask

Traditionally, if we have the ground truth depth map, the occlusion area can be obtained by comparing the difference of the warped result and the original image. However, there are some limitations to finding a good occlusion area. One problem comes from the accuracy of the predicted depth. Another problem comes from the different light conditions of different camera views. This problem is called *brightness discrepancies*. As the predicted depth is not fully accurate, the warped image result is also not fully accurate. That is why the traditional reconstruction method can not obtain accurate occlusion areas.

Instead, we propose a new occlusion mask, namely *stereo-temporal consistency (STC) occlusion mask*. The computation steps of the STC mask are shown in Fig. 2.

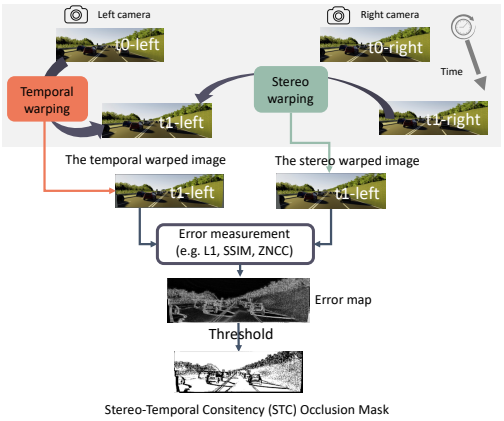


Fig. 2. The steps for computing the STC occlusion mask.

The STC mask computes the consistency between stereo and temporal warped results, it defines that different pixel intensities between them are possible stereo or temporal occlusion areas. Meanwhile, the inaccurate predicted depth values will also be reduced, because incorrect disparity values can not satisfy the temporal and stereo consistency. If the results of the stereo warping and temporal warping are different, there are only three possible reasons: (1) the temporal views occlusion areas, (2) the stereo views occlusion areas, (3) the wrong depth predictions.

To begin, the STC mask computes the stereo warped image I_{sw} and temporal warped image I_{tw} on the same stereo-temporal view, such as the left camera view at time 1 in Fig. 2. Then, the error map between the I_{sw} and the I_{tw} is computed. Finally, setting a threshold value on this error map yields an STC mask of $0/1$.

In practice, to separate the stereo occlusion and the temporal occlusion, we use the overlap set between the traditional occlusion mask and the STC mask as the final occlusion mask. There is a stereo occlusion mask (the STC-s mask) and a temporal occlusion mask (the STC-t mask). Then the STC-s mask is used in stereo image warping loss, and the STC-t mask is used in temporal image warping loss. For example, the pose estimation of adjacent frames will use the STC-t mask. Therefore, the actual percent of masked pixels is less than the given percent in the STC mask setting.

1) ZNCC measurement for STC mask: Commonly, L1 or L2 error is used to compute the error map between the warped and ground truth image. However, we find that L1 error is not a good choice because of the brightness discrepancy problem of different camera views.

To alleviate this problem, a zero-normalized cross-correlation (ZNCC) measurement is introduced into the STC mask. In theory, ZNCC measures the similarity of each $h \times w$ local patch, not only the intensity of the single pixel like the L1 error. The error measurement in this paper is (1-ZNCC), which has a range of $[0,2]$. ZNCC values on each pixel are computed with respect to their surrounding local image patch. This property promises that it is more robust for the brightness discrepancies of two views and the noises introduced by inaccurate warping.

The equation of the ZNCC is shown in Eq. 1.

$$ZNCC = \frac{1}{C}\sum_{c=0}^{C}\frac{1}{H*W}\sum_{j=0}^{H}\sum_{i=0}^{W}\left(\frac{\frac{1}{h*w}\Sigma_{y=j}^{h}\Sigma_{x=i}^{w}(F_{(y,x,c)}-\overline{F}_{(y,x,c)})*(T_{(y,x,c)}-\overline{T}_{(y,x,c)})}{\frac{1}{h*w}\Sigma_{y=j}^{h}\Sigma_{x=i}^{w}\sqrt{(F_{(y,x,c)}-\overline{F}_{(y,x,c)})^2}*\sqrt{(T_{(y,x,c)}-\overline{T}_{(y,x,c)})^2}}\right) \quad (1)$$

Where The F, T are the target image and template image separately. The H, W, and C represent the size and channel of the global image, respectively; the h, w represents the given local image patch size.

C. Local Average Max (LAM) homogeneous texture Mask

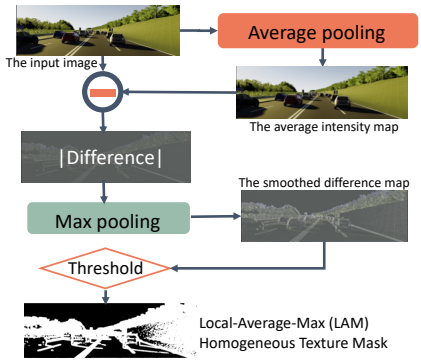


Fig. 3. The steps for computing the LAM homogeneous texture mask.

The homogeneous texture areas in the images are also hallucinated for the image warping loss optimization, because there is no significant landmark to be matched between the reconstructed and original image pair, and the depth is hard to be estimated visually.

To filter the homogeneous texture pixels, we propose a new module to generate the homogeneous texture mask, namely the local average max (LAM) homogeneous texture mask, as shown in Fig. 3.

The LAM mask proposes a new way to find homogeneous texture areas. Normally, homogeneous texture areas have similar intensities for each pixel inside. If a pixel has the same or similar intensity to the average intensity of its surrounding local image patch, this pixel should belong to a homogeneous texture area. Therefore, the LAM mask can find homogeneous texture pixels.

The steps for generating the LAM mask are as follows. Firstly, for each pixel position, we compute the average intensity of each surrounding local image patch. The average pooling layer, which is widely used in deep learning methods, is used to compute the average intensity map efficiently.

Then, the difference map between the input image and the average intensity map is obtained with the L1 measurement.

But the difference map is noisy at this step. A max pooling layer is further applied to smooth the difference map. The max pooling layer keeps the max value of the local image patch on each pixel position.

Finally, the LAM mask is also obtained by setting a threshold. The threshold can be fixed for each video sequence, or it can also be dynamic for each frame of a video sequence but the percent of masked pixels is controlled.

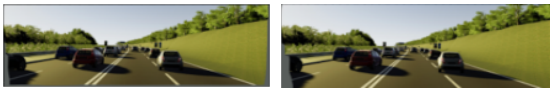


Fig. 4. The gray ares are the unrecoverable edge areas from the temporal and stereo reconstruction. The left one is the temporal reconstruction image, the right one is the stereo reconstruction image.

D. Edge Mask

The edge mask is the simple and common strategy to reduce the noises in the image warping [11]. As shown in Fig. 4, for calibrated stereo or temporal paired images, there are always some edge areas that can not be recovered from the source image view by the warping operation.

The above Fig. 4 shows two examples of the warped results on the left camera view. The left one is the temporal warped image, the right one is stereo warped image. The gray areas are the missing pixels. The stereo warped image can not recover the left edge from the right image, the temporal warped image can not recover surrounding edge areas.

III. EXPERIMENTS

A. Datasets and Implement Details

Five datasets are used for the abundant experiments in this paper. There are three datasets for the ground vehicle: KITTI Odometry, KITTI Depth, and Virtual KITTI2, and two datasets for the drone: MidAir and EuRoC MAV.

As one of the most popular benchmarks for autonomous driving. KITTI data is used for comparing our method with the state-of-the-art methods. Other datasets are used to show the significant improvement with the multi-mask system.

1) KITTI Odometry: Nine sequences (00-08) are used for training, and two sequences (09, 10) are used for testing, as in most papers. For the depth estimation network, the network is optimized with 5 training epochs on this dataset.
2) KITTI Depth: Similarly to the previous work [6], we reorganize this dataset by locating the possible ± 1 temporal adjacent images in the raw data, allowing us to perform temporal image warping and reconstruction. Following previous works, the Eigen split and evaluation method are used [1].
3) Virtual KITTI2: This is a common virtual dataset similar to the KITTI dataset. Scene{01,02,06,18} are used for the training of the depth estimation network, Scene20 is used for the evaluation. Only the data with 'morning' weather is used. The depth estimation network is trained over 30 epochs.
4) MidAir: This is a drone simulation dataset. We use all 30 sequences of 'sunny' weather for the training of the depth network, and 3 VO testing sequences are used for evaluation. Because the test sequences contain $14990, 14990, 7867$ frames, and the trajectories are repeated patterns, we only keep the first $1499, 1499, 787$ frames for the evaluation.
5) EuRoC MAV: This is also a widely used drone dataset. Following previous works, MH_03_m MH_05_d $V1\_03\_d$ $V2\_02\_m$ are used for the evaluation.

B. Experiments for the multi-mask system

To demonstrate the advantage of the multi-mask system, we show both the depth estimation and visual odometry results on both real and virtual data.

[1]Thanks for the evaluation code provided by [6]

Data Type	Mask type	Depth Error Metrics				Depth Accuracy Metric			Camera Pose Error Metric			
		abs rel	rel sqr	rmse	rmse log	$\tau < 1.25$	$< 1.25^2$	$< 1.25^3$	t_{err}(%)	r_{err} (deg/100m)	RPE_{tran} (m)	RPE_{rot} (deg)
Real seq09	Baseline	0.2827	10.4134	8.414	0.369	86.89	91.53	94.03	2.77	0.68	0.024	0.039
	LAM mask	0.2535	8.2864	7.538	0.347	86.91	91.69,	94.26	2.41	0.71	0.024	0.037
	STC mask	0.1042	1.2005	3.934	0.200	91.29	95.43	97.33	2.68	0.81	0.024	0.038
	multi-mask	**0.0665**	**0.4731**	**3.305**	**0.146**	**93.37**	**97.20**	**98.65**	**1.54**	**0.45**	**0.021**	**0.032**
Real seq10	Baseline	0.4019	14.8759	9.528	0.456	82.49	88.11	91.38	1.89	0.56	0.016	0.042
	LAM mask	0.3388	11.1628	8.387	0.418	83.18	88.80	92.13	1.52	0.47	0.016	0.040
	STC occlu	0.1587	1.9037	4.051	0.273	87.23	92.33	95.13	1.66	0.71	0.016	0.040
	multi-mask	**0.0840**	**0.4262**	**2.804**	**0.175**	**91.13**	**95.74**	**97.91**	**1.48**	**0.32**	**0.015**	**0.038**
Virtual seq20	Baseline	0.7648	84.2296	62.1419	0.6397	78.75	84.38	87.56	13.90	3.75	0.066	0.061
	LAM mask	0.3273	34.7354	63.2385	0.5206	79.81	86.98	90.35	5.29	0.87	0.045	0.017
	STC mask	0.3298	19.9154	64.7536	0.4740	79.20	86.60	90.29	1.68	0.76	0.006	0.014
	multi-mask	**0.1284**	**8.8084**	**59.4475**	**0.3545**	**83.12**	**91.16**	**94.67**	**0.95**	**0.46**	**0.005**	**0.013**

TABLE I

EXPERIMENTS TO SHOW THE EFFECT OF THE PROPOSED MULTIPLE MASKS. THREE GROUPS OF RESULTS ON BOTH REAL-WORLD DATA AND VIRTUAL DATA ARE RECORDED, AND ALL SHOW SIGNIFICANT IMPROVEMENTS WITH THE MASKS.

In our experiments, the STC mask and the LAM mask both improve depth prediction accuracy and camera pose prediction accuracy. The improvement is more significant on the simulation data, Virtual KITTI2. And the STC mask has a better effect than the LAM mask, which also suggests that removing the occlusion areas is more important.

The multi-mask system is obtained by combining the STC mask and the LAM mask. For these three sequences, the multi-mask system has the best result. The results on the depth estimation task and the visual odometry task all have significant improvements. These results not only suggest that both the occlusion areas and homogeneous texture areas are important for the computation of the image warping loss, but also demonstrate the effect of the proposed multi-mask system. The above experiment data is reported in Tab. I.

Besides the data in Tab. I, we also show the visualization of the estimated trajectory of the visual odometry. The Fig. 5 shows the comparison of the localization results with different masking strategies. This visual comparison clearly suggests that the proposed masking strategy is efficient and essential for visual odometry.

1) Ablation study: LAM mask for the visual odometry: Tab. II displays the visual odometry results with varying percentages of masked pixels from the homogeneous texture. There is a slight improvement with the homogeneous texture mask. This suggests that the homogeneous texture areas affect the accuracy of the dense direct method. And the VO results with the LAM mask are not sensitive to the setting parameter (i.e. the masked percent), which is also an advantage of the LAM mask.

%	t_{err}	r_{err}	%	t_{err}	r_{err}
1	1.89	0.71	30	1.87	0.70
5	**1.86**	**0.70**	40	1.87	0.70
10	**1.86**	**0.70**	50	1.88	0.70
20	**1.86**	**0.70**	60	1.88	0.70

TABLE II

THE VO RESULTS USING THE LAM MASK WITH VARYING PERCENTAGES OF MASKED PIXELS. THE RESULTS ARE RECORDED WITH KITTI ODOMETRY ERROR METRICS ON KITTI SEQUENCE 09.

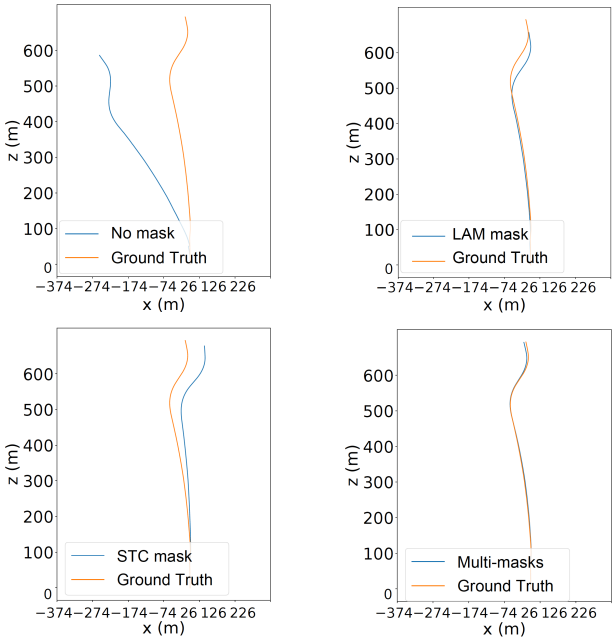


Fig. 5. The comparison of the estimated and actual visual odometry trajectory. The results of four different conditions (no mask, LAM mask, STC mask, and multi-mask) are shown. A $x-z$ dimension view (bird's eye view), is shown.

Type	t_{err}	r_{err}	Type	t_{err}	r_{err}
Base	3.74	1.21	Base	4.77	1.54
L1	3.47	1.15	L1	4.07	1.49
L1&SSIM	3.26	0.83	L1&SSIM	4.07	1.48
1-ZNCC	**1.81**	**0.68**	1-ZNCC	**3.46**	**1.30**

TABLE III

THE VO RESULTS ARE COMPARED USING DIFFERENT ERROR MEASUREMENTS IN THE STC MASK. THESE TWO TABLES ARE THE RESULTS OF KITTI SEQ 09 AND VKITTI2 SEQ 20. 'BASE' IS THE OCCLUSION MASK OBTAINED BY COMPARING THE ERRORS BETWEEN THE WARPED IMAGE AND THE ORIGINAL IMAGE.

2) Ablation study: STC mask for the visual odometry: Tab. III first shows the different VO results with different error measurement in the STC mask. Both results on two

Size	t_{err}	r_{err}	%	t_{err}	r_{err}	%	t_{err}	r_{err}
7×7	2.19	0.79	1	1.88	0.70	60	1.58	0.57
15×15	1.84	0.68	10	1.82	0.68	70	1.46	**0.50**
21×21	**1.81**	**0.68**	20	1.81	0.66	75	**1.42**	0.53
25×25	1.82	0.68	30	1.80	0.65	77.5	4.11	4.34
31×31	1.82	0.69	50	1.71	0.62	80	45.89	11.05

TABLE IV

THE VO RESULTS USING STC MASK WITH DIFFERENT ZNCC LOCAL PATCH SIZES, AS WELL AS WITH DIFFERENT PERCENT OF MASKED PIXELS, ARE SHOWN. THE RESULTS ARE RECORDED WITH KITTI ODOMETRY ERROR METRICS ON KITTI SEQUENCE 09.

different datasets suggest that the (1-ZNCC), which considers local areas on each pixel, is a better error measurement, and the L1 loss, which only considers single intensity on each pixel, is not a good choice. And the comparison between the STC mask and the traditional baseline occlusion mask suggests that the proposed method is better for finding the possible occlusion areas.

For the details of the (1-ZNCC) measurement, Tab. IV shows the results of different local patch sizes in the ZNCC and the results with different percents of masked pixels. As mentioned in Sec. II-B, the actual masked pixels percent is less than the given percent. The STC mask requires the local patch size in the ZNCC to be large enough (i.e., $>=21\times21$) to maintain good VO performance. And the VO performance is stable when the local patch size is large enough. But the computation cost also increases with the increase of the local patch size. $21-25$ pixels local patch size is a suitable choice for the ZNCC measurement in the STC mask. For the percent of masked pixels, the best result is obtained when setting masked pixels to 75% (about 50% of actual masked pixels). After that, the odometry error increases quickly.

C. Compare with the State-of-the-art Methods

To show the advantage of the proposed masks, we compare the proposed masked DDM visual odometry method with the current state-of-the-art algorithms on depth estimation and visual odometry separately. Tab. V shows that the multi-mask system can help the baseline depth estimation network [2] achieve new state-of-the-art depth results on KITTI Depth (eigen split) dataset. This is the most popular benchmark for this task.

Tab. VI shows the localization results using the KITTI odometry metrics [13]. These results all suggest that the traditional dense direct method (DDM) with the proposed multi-mask system can achieve competitive performance with the state-of-the-art methods, and it realizes new SOTA results on sequence 10.

Tab. VII shows the localization results using ATE metric [7]. Compared with those works recording the ATE metric, the proposed multi-mask system helps DDM to realize new state-of-the-art results.

All of these above results can demonstrate the advantage of the proposed multi-mask system. It helps the baseline method achieve state-of-the-art performance.

Fig. 6 shows the estimated trajectory with our method.

year	Method	Error Metric				Accuracy Metric		
		abs rel	rel sqr	rmse	rmse log	1.25	1.25^2	1.25^3
2017	[7]	0.208	1.768	6.856	0.283	67.8	88.5	95.7
2018	[14]	0.149	1.060	5.567	0.226	79.6	93.5	97.5
2019	[15]	0.128	0.935	5.011	0.209	83.1	94.5	97.9
2019	[6]	0.106	0.806	4.630	0.193	87.6	95.8	**98.0**
2020	[3]	0.099	0.763	4.485	0.185	88.5	95.8	97.9
2021	[16]	0.121	0.971	5.206	0.214	84.3	94.4	97.5
2021	[17]	0.105	0.842	4.810	0.196	86.1	94.7	97.8
2022	[2]	0.080	0.795	4.146	0.185	92.2	95.9	97.6
	Ours	**0.077**	**0.676**	**3.863**	**0.173**	**92.7**	**96.3**	97.9

TABLE V

COMPARE THE MULTI-MASK-BASED NETWORK WITH STATE-OF-THE-ART DEPTH ESTIMATION NETWORKS TRAINED WITHOUT GROUND TRUTH DEPTH ANNOTATIONS.

Year	Method	seq.9		seq.10	
		t_{err}	r_{err}	t_{err}	r_{err}
	model-based				
2015	ORB [18]	15.30	0.26	3.68	0.48
	end-to-end				
2017	SfmLearner [7]	17.84	6.78	37.91	17.80
2018	Geonet [14]	43.76	16.00	35.60	13.8
2019	Wang [19]	9.30	3.50	7.21	3.90
2019	Li [20]	8.10	2.81	12.90	3.17
2021	TAPE [21]	6.72	2.60	8.66	3.13
2021	F2FPE [21]	2.36	1.06	3.00	1.28
	hybrid				
2018	DVSO [22]	0.83	0.21	0.74	0.21
2019	UnOS [23]	5.21	1.80	5.20	2.18
2020	D3VO [3]	0.78	×	0.62	×
2020	DFVO [24]	2.07	0.23	2.06	0.36
2022	PDENet-DPE [2]	0.87	0.28	0.87	0.46
	Ours	**0.76**	0.41	**0.42**	0.24

TABLE VI

THE SOTA RESULTS WITH KITTI METRICS t_{err}, r_{err} ON SEQ 09, 10.

Year	Method	seq.9	seq.10
		ATE	*ATE*
	model-based		
2015	ORB full [18]	0.0140±0.0080	0.0120±0.0110
2015	ORB short [18]	0.0640±0.1410	0.0640±0.1300
	end-to-end		
2017	SfmLearner [7]	0.0160±0.0090	0.0130±0.0090
2018	Geonet [14]	0.0120±0.0070	0.0120±0.0090
2018	Vid2depth [25]	0.0130±0.0100	0.0120±0.0110
2019	Com Col [26]	0.0120±0.0070	0.0120±0.0080
	hybrid		
2019	UnOS [23]	0.0120±0.0060	0.0130±0.0080
2022	PDENet-DPE [2]	0.0109±0.0068	0.0105±0.0088
	Ours	**0.0109±0.0064**	**0.0099±0.0089**

TABLE VII

THE SOTA RESULTS WITH *ATE* METRIC ON KITTI SEQ 09, 10.

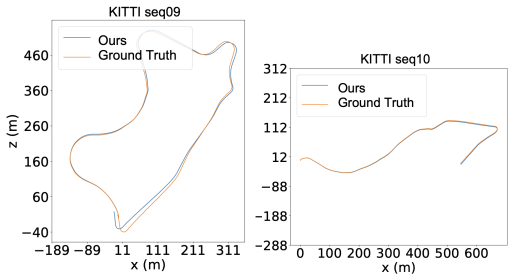
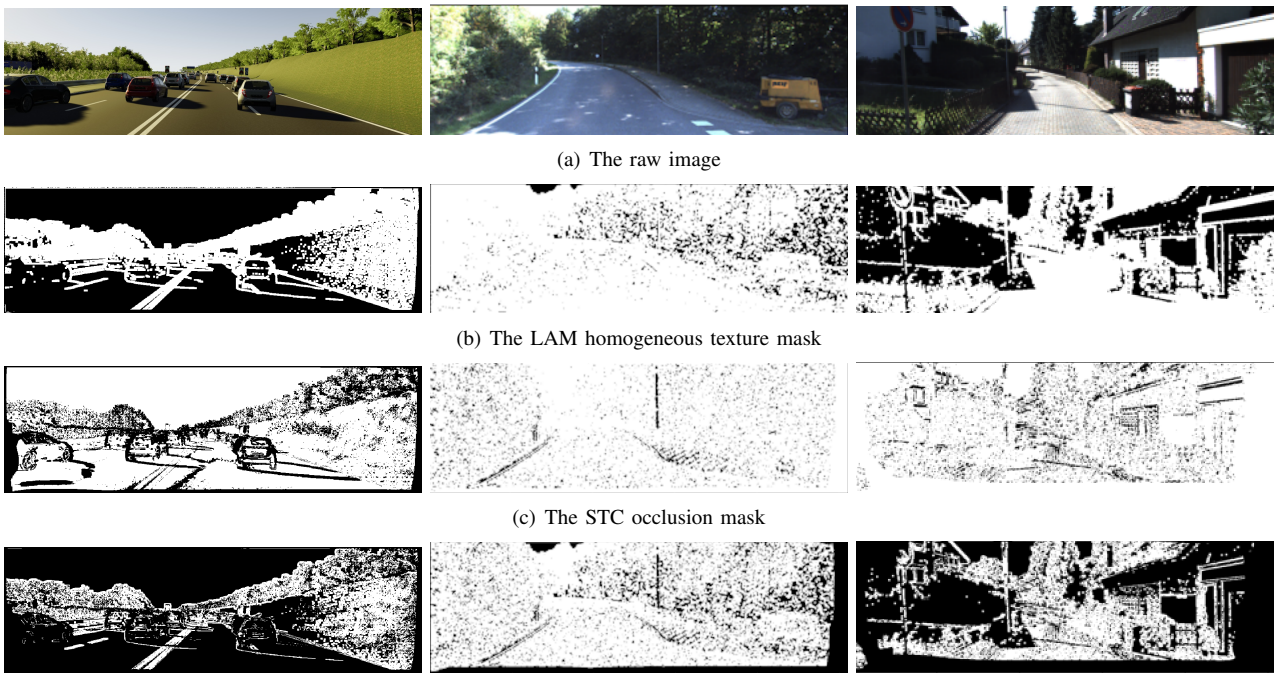


Fig. 6. The estimated trajectories on KITTI sequence 09 and 10.



(a) The raw image

(b) The LAM homogeneous texture mask

(c) The STC occlusion mask

(d) The multi-mask

Fig. 7. The masks examples from sequence 20 of the Virtual KITTI2 dataset and sequence 09 and 10 of the KITTI Odometry dataset (from the left to the right). The black areas are the masked pixels. The LAM homogeneous texture mask removes homogeneous texture areas, e.g. the sky, highlight and shadow areas. The STC occlusion mask removes occlusion areas. The multi-mask system is obtained by removing all masked pixels from these masks.

D. More Comparisons in Different Scenes

In this part, we conduct experiments to demonstrate that the DDM with the proposed multi-mask system not only works well on vehicle-captured datasets, but can also improve the visual odometry accuracy for drone-captured videos.

Seq ID	*MH*_03	*MH*_05	*V*1_03	*V*2_02	Mean
Baseline	0.0061	0.0049	0.0095	0.0093	0.0075
Multi-mask	**0.0052**	**0.0029**	**0.0049**	**0.0078**	**0.0052**

TABLE VIII

THE VO RESULTS WITH ATE METRIC ON THE EUROC MAV DATASET.

Seq ID	Sunny00	Sunny01	Sunny02	Mean
Baseline	0.1069	0.1375	0.0757	0.1067
Multi-mask	**0.0082**	**0.0164**	**0.0073**	**0.0319**

TABLE IX

THE VO RESULTS WITH ATE METRIC ON THE MID-AIR DATASET.

Tab. VIII and Tab. IX show the visual odometry results on the EuRoC MAC and MidAir drone datasets separately. In the MidAir dataset, some frames have severe occlusion problems when the drone is close to the forest or the hill in the MidAir dataset. And there are larger homogeneous texture areas (e.g. the sky, the lake) in this dataset. Therefore, the improvement with the multi-mask system is more significant in this dataset.

E. Visualize the multi-mask system

Fig. 7 shows some examples of the proposed masks on two datasets. For the homogeneous texture mask, the sky, highlight and shadow areas are masked, while the significant areas, like the traffic lane, are kept. For the occlusion mask, the most possible occlusion areas, like the edge areas of objects, are successfully masked.

IV. CONCLUSION

In this paper, we proposed a multi-mask approach to increase the accuracy in the computation of the image warping loss in the presence of occlusions and homogeneous texture areas. The multi-mask approach can be applied both to the unsupervised depth estimation network and the DDM-based visual odometry and is mainly composed by a STC mask and a LAM mask. The STC mask takes the advantage of the spatial-temporal relations between successive images to find occlusion areas. The LAM mask uses the relations inside the local image patch to find the homogeneous areas automatically with only RGB intensities. In the future, this multi-mask system will further include semantics-based masks for dynamic objects. This will further improve the robustness of the dense direct visual odometry methods.

ACKNOWLEDGMENTS

This work was funded by 3IA institute at Université Côte d'Azur. The results have been obtained using OPAL computing cluster of INRIA and Université Côte d'Azur.

REFERENCES

[1] A. I. Comport, E. Malis, and P. Rives, "Real-time quadrifocal visual odometry," *IJRR*, vol. 29, no. 2-3, pp. 245–266, 2010.

[2] Z. Liu, E. Malis, and P. Martinet, "A new dense hybrid stereo visual odometry approach," in *IROS*, 2022.

[3] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *CVPR*. IEEE, 2020, pp. 1281–1292.

[4] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos," in *CVPR*, 2019, pp. 8071–8081.

[5] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*. IEEE, 2017, pp. 270–279.

[6] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*. IEEE, 2019, pp. 3828–3838.

[7] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*. IEEE, 2017, pp. 1851–1858.

[8] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *CVPR*. IEEE, 2018, pp. 2022–2030.

[9] G. Wang, H. Wang, Y. Liu, and W. Chen, "Unsupervised learning of monocular depth and ego-motion using multiple masks," in *ICRA*. IEEE, 2019, pp. 4724–4730.

[10] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *NeurIPS*, vol. 32, 2019.

[11] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *CVPR*, 2018, pp. 5667–5675.

[12] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia, "Unsupervised learning of geometry with edge-aware depth-normal consistency," in *AAAI*, 2018.

[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.

[14] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *CVPR*. IEEE, 2018, pp. 1983–1992.

[15] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2624–2641, 2019.

[16] C. Ling, X. Zhang, and H. Chen, "Unsupervised monocular depth estimation using attention and multi-warp reconstruction," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[17] X. Ye, X. Fan, M. Zhang, R. Xu, and W. Zhong, "Unsupervised monocular depth estimation via recursive stereo distillation," *IEEE Transactions on Image Processing*, vol. 30, pp. 4492–4504, 2021.

[18] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *TRO*, vol. 31, no. 5, pp. 1147–1163, 2015.

[19] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth," in *CVPR*. IEEE, 2019, pp. 5555–5564.

[20] Y. Li, Y. Ushiku, and T. Harada, "Pose graph optimization for unsupervised monocular visual odometry," in *ICRA*. IEEE, 2019, pp. 5439–5445.

[21] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Transformer guided geometry model for flow-based unsupervised visual odometry," *Neural Computing and Applications*, pp. 1–12, 2021.

[22] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *ECCV*. Springer, 2018, pp. 817–833.

[23] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos," in *CVPR*. IEEE, 2019, pp. 8071–8081.

[24] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *ICRA*. IEEE, 2020, pp. 4203–4210.

[25] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *CVPR*. IEEE, 2018, pp. 5667–5675.

[26] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *CVPR*. IEEE, 2019, pp. 12 240–12 249.